PATENT ABSTRACTS OF JAPAN

(11)Publication number : 09-258907
(43)Date of publication of application : 03.10.1997

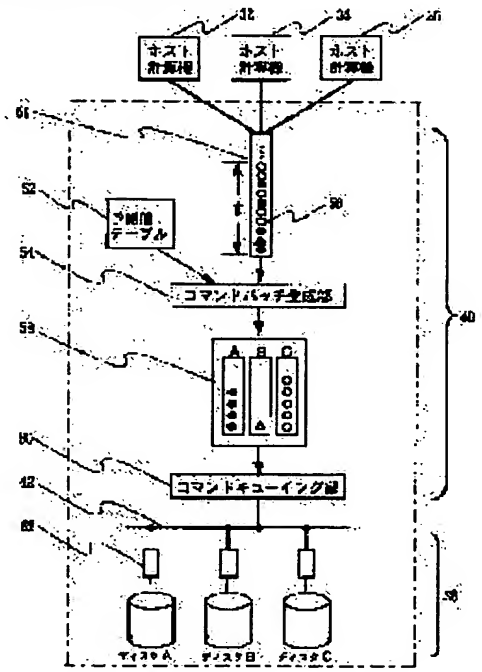
(51)Int.Cl. G06F 3/06
G06F 3/06
G06F 13/10
G11B 19/02

(21)Application number : 08-068405
(22)Date of filing : 25.03.1996

(71)Applicant : MITSUBISHI ELECTRIC CORP
(72)Inventor : NAGASHIMA MASARU

(54) HIGHLY AVAILABLE EXTERNAL STORAGE DEVICE HAVING PLURAL STORAGE DISK PARTS

(57)Abstract:
PROBLEM TO BE SOLVED: To improve a response and throughput ad to eliminate the imbalance of the response and throughput between storage disk parts by most suitably processing the queue of a command issued to an external; storage device.
SOLUTION: The command batch generation part 54 of the external storage control part 40 reads the number of commands whose sum of processing time prediction values by a prediction value table 52 becomes prescribed time as a group (batch) from a reception queue 50. The batch is stored in a transmission quene 58. A command quauing part 60 repeats a processing for selecting a disk device 38 whose prediction value of command processing time is the longest, for taking out the command for it and issuing it to the disk device 38. When the transmission queue 58 becomes vacant, the next batch is read into it.



LEGAL STATUS

[Date of request for examination]

[Date of sending the examiner's decision of rejection]

[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]

[Date of final disposal for application]

[Patent number]

[Date of registration]

[Number of appeal against examiner's decision of rejection]

[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁（ＪＰ）　(12) 公開特許公報（Ａ）　(11)特許出願公開番号

特開平9－258907

(43)公開日　平成９年(1997)10月３日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ	3/06	３０１		Ｇ０６Ｆ	3/06	３０１Ｆ
		５４０				５４０
	13/10	３４０			13/10	３４０Ｂ
Ｇ１１Ｂ	19/02	５０１		Ｇ１１Ｂ	19/02	５０１Ｆ

審査請求　未請求　請求項の数10　ＯＬ　（全 16 頁）

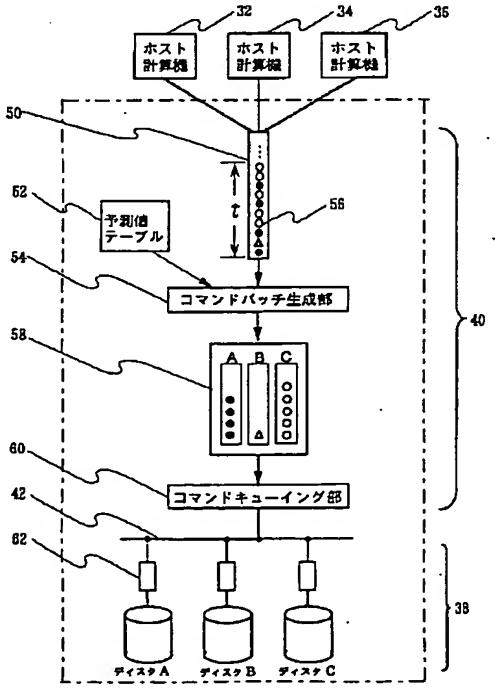
(21)出願番号　特願平8－68405

(22)出願日　平成８年(1996)３月25日

(71)出願人　000006013
三菱電機株式会社
東京都千代田区丸の内二丁目２番３号
(72)発明者　長島　勝
東京都千代田区丸の内二丁目２番３号　三菱電機株式会社内
(74)代理人　弁理士　吉田　研二　（外２名）

(54)【発明の名称】　複数の記憶ディスク部を有した高可用性の外部記憶装置

(57)【要約】
【課題】　外部記憶装置に対して発行されたコマンドのキューが最適に処理されておらず、レスポンスやスループットといった性能が不十分である。
【解決手段】　外部記憶制御部４０のコマンドバッチ生成部５４は、予測値テーブル５２によるコマンドの処理時間予測値の和が所定時間となる数のコマンドをグループ（バッチ）として受入キュー５０から読み出す。バッチは送出キュー５８に格納される。コマンドキューイング部６０はその中のコマンド処理時間の予測値が最長のディスク装置３８を選択し、それに対するコマンドを取り出して対応するディスク装置３８に発行するという動作を繰り返す。送出キュー５８は空になると、これに次のバッチが読み込まれる。



1

【特許請求の範囲】

【請求項１】　データを格納しこのデータに対する読み出し／書き込みコマンドをキューイング可能な複数の記憶ディスク部と、コンピュータからの前記コマンドを受入キューに蓄え前記記憶ディスク部に順次発行し、前記コンピュータと前記記憶ディスク部との間のデータ入出力処理を制御する外部記憶制御部とを備えた外部記憶装置において、

前記外部記憶制御部は、

前記コマンドの処理に要すると予測される時間であるコマンド処理時間予測値を生成する予測処理時間生成手段と、

所定のタイミングで前記受入キュー内のコマンドをその前記コマンド処理時間予測値の和が所定の処理タイムスライスに応じた値となる個数だけ格納される、前記コマンドの待ち行列である送出キューと、

前記予測処理時間生成手段に基づき前記各記憶ディスク部別の処理時間を予測する処理時間予測手段と、

この予測された処理時間が最大の前記記憶ディスク部に対する前記コマンドを前記送出キューから取り出して、対応する前記記憶ディスク部にキューイングするコマンドキューイング手段と、

前記各送出キューが空になると、前記受入キューから前記送出キューへ前記コマンドを格納するコマンドバッチ生成手段と、

を含むことを特徴とする外部記憶装置。

【請求項２】　請求項１記載の外部記憶装置において、

前記外部記憶制御部と前記記憶ディスク部はバスにより接続され、

前記外部記憶制御部は、前記記憶ディスク部に前記コマンドを試行的に発行し、そのコマンドに対し前記記憶ディスク部が前記バスの使用権を獲得するまでの応答時間を計測する応答計測手段を有し、

前記コマンドキューイング手段は、前記コマンドの発行を前記バスのフリー状態の検出から前記応答時間に応じた時間、遅延させる遅延手段を有すること、

を特徴とする外部記憶装置。

【請求項３】　請求項１記載の外部記憶装置において、

前記外部記憶制御部は、

前記記憶ディスク部に対する読み出しコマンドの処理待ち数を計数するコマンド計数手段と、

複数の読み出しコマンドに対する前記記憶ディスク部からの読み出しデータを保持するバッファを有し、

前記処理待ち数に基づいて前記バッファを制御し、複数の読み出しコマンドの前記読み出しデータを前記バッファへ蓄積させ、その蓄積された読み出しデータをコンピュータへ一括転送させるバッファ制御手段と、

を含むことを特徴とする外部記憶装置。

【請求項４】　請求項１記載の外部記憶装置において、

前記外部記憶制御部は、

2

前記各送出キューから送出され処理待ち状態にあるコマンドについての前記コマンド処理時間予測値の最大値をタイムアウト値として前記各記憶ディスク部ごとに保持するタイムアウト値保持手段と、

直前のコマンドの処理終了からの経過時間を、前記各記憶ディスク部ごとに計測する経過時間計測手段と、

前記経過時間がその対応する前記記憶ディスク部の前記タイムアウト値を越えたときエラーとして判定するエラー判定手段と、

を含むことを特徴とする外部記憶装置。

【請求項５】　請求項１記載の外部記憶装置において、

前記コマンドキューイング手段は、前記コマンドによりアクセスされる前記記憶ディスク部のアドレスに基づいて、アクセス時間が最小と予測されるコマンドを選択するアクセス最適化手段を有することを特徴とする外部記憶装置。

【請求項６】　請求項５記載の外部記憶装置において、

前記外部記憶制御部が複数個設けられて前記記憶ディスク部を共有し、

この各外部記憶制御部は、アクセスした前記記憶ディスク部のアドレスに基づくアドレス情報を他の外部記憶制御部に通知するアドレス情報通知手段を有し、

前記アクセス最適化手段は、前記アドレス情報を参照して前記送出キューから前記コマンドを選択すること、

を特徴とする外部記憶装置。

【請求項７】　請求項１記載の外部記憶装置において、

前記記憶ディスク部は、２つのディスク装置を含み、これらディスク装置間においてミラーリングが行われるレベル１のＲＡＩＤアレイ・ディスクであり、

前記外部記憶制御部は、前記送出キューの中の前記２つのディスク装置に対する読み出しコマンドを、それらの前記コマンド処理時間予測値に基づいて前記２つのディスク装置に振り分けて発行し、これら両ディスク装置の読み出し処理時間の均等化を図るＲＡＩＤ１負荷分散手段を有すること、

を特徴とする外部記憶装置。

【請求項８】　請求項１記載の外部記憶装置において、

前記記憶ディスク部は、複数のディスク装置を含み、これら各ディスク装置に格納されるデータに基づいて生成されるパリティをこれら全ディスク装置に分割して保持するレベル５のＲＡＩＤディスク・アレイであり、

前記外部記憶制御部は、書き込みコマンドの処理における新しいパリティの生成処理に用いるデータとして、更新されるデータと現パリティとを読み出すか、これら以外の現パリティの生成に用いたデータを読み出すかを、読み出し処理時間が前記ディスク装置間にて均等化されるように前記コマンド処理時間予測値に基づいて選択するＲＡＩＤ５負荷分散手段を有すること、

を特徴とする外部記憶装置。

【請求項９】　請求項１記載の外部記憶装置において、

前記記憶ディスク部は、複数のディスク装置を含み、前記各ディスク装置に格納されるデータに基づいてそのセクタ単位にパリティを生成し、このパリティを特定の前記ディスク装置に集中して保持させて高転送速度を実現するレベル３のＲＡＩＤディスク・アレイであり、

前記外部記憶制御部は、分割された複数の前記セクタ範囲ごとに、パリティを保持する前記ディスク装置が異なるようにパリティ格納位置を生成するパリティ位置割付手段を有すること、

を特徴とする外部記憶装置。

【請求項１０】　請求項７から請求項９までのいずれかに記載の外部記憶装置において、

前記外部記憶制御部は、起動時に前記ディスク装置のシーク時間及び回転時間を測定し、この測定値に基づいて前記コマンド処理時間予測値を設定する予測処理時間設定手段を有することを特徴とする外部記憶装置。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、計算機システムに用いられる複数の記憶ディスク部を有した外部記憶装置に関し、特にそのレスポンスとスループットの向上による可用性の向上に関する。

【０００２】

【従来の技術】ディスク装置における読み出し／書き込み処理には機械的動作が伴うため、その処理速度は、計算機のＣＰＵ処理速度に比べると段違いに遅い。しかし、ディスク装置は、大量のデータを安価に記憶できる点で優れている。このため、現在の計算機システムにおいては、上記処理速度の格差を補う様々な工夫をした上で、ディスク装置は大規模容量を有する外部記憶装置に採用されている。

【０００３】図１５は、例えば特開平６－１９６２５に示される従来の外部記憶装置の模式的な構成図である。この外部記憶装置２は、２つの記憶ディスク部４、６を含み、２つのディスク制御部８、１０がディスク接続部１２を介して記憶ディスク部４、６に接続されている。ディスク制御部８、１０にはそれぞれホスト計算機１４、１６、１８に接続され、各ホストコンピュータは一方のディスク制御部が使用中でももう一方のディスク制御部にアクセスし外部記憶装置２を利用することができる。各ディスク制御部はホスト計算機１４、１６、１８が発行する外部記憶装置２に対するコマンドに基づいて、各記憶ディスク部を制御する。

【０００４】ホスト計算機１４、１６、１８が発行するコマンドはディスク接続部１２内の共有メモリ２０に設けられる待ち行列（キュー）に格納される。キュー内のコマンドはバスを介して順次、記憶ディスク部に対し送られる。記憶ディスク部がコマンドのキューイング機能を有する場合、すなわち記憶ディスク部自身が受け付けたコマンドのキューを備えている場合には、共有メモリ

内の前記キューからのコマンドの送出は、それ以前に送出されたコマンドの処理終了を待たずに行うことができる。

【０００５】上記のようにディスク制御部を複数にすることにより、ディスク制御部の獲得におけるホストコンピュータ間の競合が緩和され、ホストコンピュータが外部記憶装置２に対してアクセスするまでの時間が短縮される。また、記憶ディスク部を複数にすることにより、１つの記憶ディスク部がヘッドのシークによるポジショニング動作やディスクの回転動作を行う間に、ディスク制御部８、１０と他の記憶ディスク部との間でコマンドの発行やデータ転送を行うことができ、コンピュータが実行するプログラム全体でのレスポンス時間の低減やスループットの向上が図られる。また、メモリ２０をディスクキャッシュとして機能させてもレスポンス時間を低減することができる。

【０００６】さて、外部記憶装置の処理性能を向上させるため前記キューを用いる方法として、特開平６－２５９１９８に、アクセスされる記憶ディスク部の領域を連続とすることができる場合には、それに応じてキュー中のコマンドを並べ換え、これにより上記ポジショニング動作やディスクの回転動作を効率化する方法が示されている。

【０００７】

【発明が解決しようとする課題】従来は、キュー内のコマンドをどのように処理すれば外部記憶装置の処理性能が向上するかについては十分な配慮がなされているとは言えず、外部記憶装置の処理はレスポンスやスループットに関し必ずしも効率的には行われていないという問題点があった。例えば、上述の構成のように複数の記憶ディスク部がバスを介してディスク制御部に接続されている場合ではバスアービトレーション、すなわちバスの使用権の獲得等を考慮しなければ、最適な処理を行うことはできない。ここで、複数の記憶ディスク部を有する外部記憶装置における最適な処理とは、装置全体のレスポンスやスループットが優れているだけでなく、記憶ディスク部間におけるレスポンスやスループットの不均衡がないことである。

【０００８】本発明は複数の記憶ディスク部を有した上記最適な処理を実現する外部記憶装置を提供することを目的とする。

【０００９】

【課題を解決するための手段】本発明に係る外部記憶装置においては、外部記憶制御部は、コマンドの処理に要すると予測される時間であるコマンド処理時間予測値を生成する予測処理時間生成手段と、所定のタイミングで受入キュー内のコマンドをその前記コマンド処理時間予測値の和が所定の処理タイムスライスに応じた値となる個数だけ格納される前記コマンドの待ち行列である送出キューと、前記予測処理時間生成手段に基づき各記憶デ

ィスク部別の処理時間を予測する処理時間予測手段と、この予測された処理時間が最大の前記記憶ディスク部に対する前記コマンドを前記送出キューから取り出して対応する前記記憶ディスク部にキューイングするコマンドキューイング手段と、前記各送出キューが空になると前記受入キューから前記送出キューへ前記コマンドを格納するコマンドバッチ生成手段と、を含むことを特徴とする。

【００１０】本発明によれば、送出キューが空になると、受入キューに蓄えられたコンピュータからのコマンドが一群ずつ取り出され、送出キューに振り分けられる。このとき、この一群をなすコマンドの個数は、そのコマンド処理時間予測値の和が所定の処理タイムスライスに応じた値となるように定められる。送出キューから各記憶ディスク部へのコマンドの発行は、キュー内のコマンドの前記コマンド処理時間予測値の和が最大である送出キューを選択し、そのキュー内のコマンドを所定個数（例えば１個）、取り出して送出するという動作を繰り返す。１回の送出動作はある程度の時間を要する。これは例えば、外部記憶制御部と記憶ディスク部とがバスで接続されている場合には、１回の送出動作に、外部記憶制御部によるバスアービトレーションの時間やバスの専有時間が必要であるためである。各記憶ディスク部はキューイング可能であり、複数のコマンドを受け付けることができる。よって基本的にはある記憶ディスク部に対する送出動作は、その記憶ディスク部で前に送出されたコマンドの処理を行っている最中でも可能である。前記コマンド処理時間予測値は、例えば記憶ディスク部の最大シーク時間や最大回転待ち時間などに基づいて予測処理時間生成手段により生成される。この予測処理時間生成手段は、コマンド処理時間予測値を格納したテーブルでもよい。さて、ヘッドのシークやディスクの回転といった機械的動作など各記憶ディスク部における処理時間の多くは記憶ディスク部間で独立であり、その部分は並列処理することができる。ここで上記のように処理時間長の長い記憶ディスク部に対するコマンドの処理を優先して開始することにより、この並列処理の割合が向上し、上記送出キューに振り分けられた一群のコマンドの各処理時間が短縮される。また、ある一群のコマンドの各記憶ディスク部への送出が終了してから、次の一群のコマンドの送出が開始される。そのためどのコマンドの処理も基本的に処理タイムスライス以下に終了し、最大レスポンス時間及び最低スループットが保証される。

【００１１】本発明に係る外部記憶装置においては、前記外部記憶制御部と前記記憶ディスク部はバスにより接続され、前記外部記憶制御部は前記記憶ディスク部に前記コマンドを試行的に発行し、そのコマンドに対し前記記憶ディスク部が前記バスの使用権を獲得するまでの応答時間を計測する応答計測手段を有し、前記コマンドキューイング手段は前記コマンドの発行を前記バスのフリー状態の検出から前記応答時間に応じた時間、遅延させる遅延手段を有すること、を特徴とする。一般に、外部記憶制御部はバス・フリー検出からアービトレーションの実行までを記憶ディスク部より短い時間で行うことができる。本発明によれば、例えば記憶ディスク部に格納されたデータに影響を及ぼさないコマンドを試行的に発行し、その記憶ディスク部からのレスポンスが外部記憶制御部に戻るまでの時間に基づいて応答時間を定める。望ましくはこの応答時間として、コマンドの試行的な発行を複数回行ってそれらの平均を採用するのがよい。この応答時間だけ外部記憶制御部から記憶ディスク部へのコマンドの発行を遅延させることにより、記憶ディスク部のバス獲得における優先度が増し、記憶ディスク部からのレスポンス時間が短縮される。

【００１２】本発明に係る外部記憶装置においては、前記外部記憶制御部は、前記記憶ディスク部に対する読み出しコマンドの処理待ち数を計数するコマンド計数手段と、複数の読み出しコマンドに対する前記記憶ディスク部からの読み出しデータを保持するバッファを有し、前記処理待ち数に基づいて前記バッファを制御し複数の読み出しコマンドの前記読み出しデータを前記バッファへ蓄積させ、その蓄積された読み出しデータをコンピュータへ一括転送させるバッファ制御手段と、を含むことを特徴とする。

【００１３】本発明によれば、コマンド計数手段が、受入キューから取り出された一群のコマンド中の読み出しコマンドのうち記憶ディスク部からレスポンスが帰ってきていないコマンド数である前記処理待ち数をカウントする。バッファ制御手段はこの処理待ち数に基づき記憶ディスク部から読み出されるデータがあるか否かを判断し、基本的にまだ読み出されるデータがある場合にはバッファからコンピュータへの転送を見合わせて次の読み出しデータの到着を待ち、一方、もう読み出されるデータがない場合やバッファの残り空き容量が少ない場合には、バッファからコンピュータへの転送を行う。このように複数の読み出しコマンドに対応するデータを一括してコンピュータへ転送することにより、その転送の際の外部記憶装置からコンピュータへのチャネルの接続処理といった時間的オーバーヘッドが低減され、レスポンスやスループットの向上が図られる。

【００１４】本発明に係る外部記憶装置においては、前記外部記憶制御部は、前記各送出キューから送出され処理待ち状態にあるコマンドについての前記コマンド処理時間予測値の最大値をタイムアウト値として前記各記憶ディスク部ごとに保持するタイムアウト値保持手段と、直前のコマンドの処理終了からの経過時間を、前記各記憶ディスク部ごとに計測する経過時間計測手段と、前記経過時間がその対応する前記記憶ディスク部の前記タイムアウト値を越えたときエラーとして判定するエラー判定手段と、を含むことを特徴とする。

【００１５】本発明によれば、前記経過時間はコマンドの処理が終了すればリセットされるので、経過時間はコマンドの処理が正常に行われているならば対応するタイムアウト値よりも小さい値である。一方、前記タイムアウト値は前記各送出キューから送出され記憶ディスク部側にキューイングされたコマンドのうち未処理のコマンドについての前記コマンド処理時間予測値の最大値であるので、経過時間がタイムアウト値より大きくなるということは、それら未処理コマンドのうちいずれのコマンドも終了しないという異常を示している。エラー判定手段は、この経過処理時間がタイムアウト値を超えたタイミングにおいてエラーを検出する。よって、エラー判定手段は、前記各送出キューの予測された処理時間の終了より早くエラーを検出することができる。

【００１６】本発明に係る外部記憶装置においては、前記コマンドキューイング手段は、前記コマンドによりアクセスされる前記記憶ディスク部のアドレスに基づいて、アクセス時間が最小と予測されるコマンドを選択するアクセス最適化手段を有することを特徴とする。

【００１７】本発明によれば、アクセス最適化手段は、前記送出キュー内の前記コマンドにより指定される前記アドレスが例えば昇順、降順となるようにコマンドを選択し、これらコマンドの実行における記憶ディスク部のヘッドのシーク時間やディスクの回転時間といったアクセス時間を最小とする。これによりレスポンスやスループットの向上が図られる。

【００１８】本発明に係る外部記憶装置においては、前記外部記憶制御部が複数個設けられて前記記憶ディスク部を共有し、この各外部記憶制御部はアクセスした前記記憶ディスク部のアドレスに基づくアドレス情報を他の外部記憶制御部に通知するアドレス情報通知手段を有し、前記アクセス最適化手段は前記アドレス情報を参照して前記送出キューから前記コマンドを選択すること、を特徴とする。

【００１９】本発明によれば、アドレス情報通知手段が、一方の外部記憶制御部において最後にアクセスされたアドレスに関する情報を、次の処理を行う外部記憶制御部に通知する。次の処理を行う外部記憶制御部はこのアドレス情報に基づいて、例えば前記アクセス時間が最小であるコマンドから処理を開始する。これによりレスポンスやスループットの向上が図られる。

【００２０】本発明に係る外部記憶装置においては、前記記憶ディスク部は２つのディスク装置を含みこれらディスク装置間においてミラーリングが行われるレベル１のＲＡＩＤアレイ・ディスクであり、前記外部記憶制御部は前記送出キューの中の前記２つのディスク装置に対する読み出しコマンドをそれらの前記コマンド処理時間予測値に基づいて前記２つのディスク装置に振り分けて発行しこれら両ディスク装置の読み出し処理時間の均等化を図るＲＡＩＤ１負荷分散手段を有すること、を特徴とする。レベル１のＲＡＩＤアレイ・ディスクでは、読み出し処理は一方のディスク装置に対してのみ行えばよい。本発明によれば、送出キュー中に複数の読み出しコマンドがあれば、それらをレベル１のＲＡＩＤアレイ・ディスクを構成する２つのディスク装置に振り分けて発行し、これら両ディスク装置において異なる読み出しコマンドを並列に処理させることにより、コマンド処理の効率が向上する。

【００２１】本発明に係る外部記憶装置においては、前記記憶ディスク部は複数のディスク装置を含みこれら各ディスク装置に格納されるデータに基づいて生成されるパリティをこれら全ディスク装置に分割して保持するレベル５のＲＡＩＤディスク・アレイであり、前記外部記憶制御部は書き込みコマンドの処理における新しいパリティの生成処理に用いるデータとして、更新されるデータと現パリティとを読み出すかこれら以外の現パリティの生成に用いたデータを読み出すかを、読み出し処理時間が前記ディスク装置間にて均等化されるように前記コマンド処理時間予測値に基づいて選択するＲＡＩＤ５負荷分散手段を有すること、を特徴とする。

【００２２】レベル５のＲＡＩＤディスク・アレイでは、データの書き込み処理はパリティの更新処理を伴う。例えばｎ個のデータをＤｋ（ｋ＝１～ｎ）、これらにより生成されるパリティをＰとする。これらはレベル５のＲＡＩＤディスク・アレイを構成する各ディスク装置に分散して保持されている。ここでＤｊ（１≦ｊ≦ｎ）をＤｊ’で更新する書き込み処理を考える。新たなｎ個のデータに対するパリティを求める第１の方法はＤｊ’、Ｄｊ及びＰから求める方法であり、第２の方法はＤｊ’及びＤｋ（ｋ≠ｊ）から求める方法である。この第１の方法においてはＤｊ及びＰをディスク装置から読み出さなければならない。一方、第２の方法においてはＤｋ（ｋ≠ｊ）を読み出さなければならない。この第１の方法と第２の方法とでは、読み出し処理が行われるディスク装置が異なる。本発明によれば、ＲＡＩＤ５負荷分散手段は送出キュー中に複数の書き込みコマンドがあれば、それらに伴う上記パリティ更新における読み出し処理が、各ディスク装置に分散し処理時間が均等化するように、上記第１の方法、第２の方法を選択する。すなわち書き込みコマンドに伴う読み出し処理が各ディスク装置間において並列に処理されることにより、コマンド処理の効率が向上する。

【００２３】本発明に係る外部記憶装置においては、前記記憶ディスク部は複数のディスク装置を含み前記各ディスク装置に格納されるデータに基づいてそのセクタ単位にパリティを生成しこのパリティを特定の前記ディスク装置に集中して保持させて高転送速度を実現するレベル３のＲＡＩＤディスク・アレイであり、前記外部記憶制御部は分割された複数の前記セクタ範囲ごとに、パリティを保持する前記ディスク装置が異なるようにパリテ

ィ格納位置を生成するパリティ位置割付手段を有すること、を特徴とする。

【００２４】本発明によれば、例えばｍ個のディスク装置から構成されるレベル３のＲＡＩＤディスク・アレイにおいて、データはｍ個のディスク装置に格納される。これにより、従来の１つのディスク装置がパリティ専用であって残りの（ｍ－１）個がデータを格納するディスク・アレイよりも、データの分散度合いが高まり、コマンド処理が各ディスク装置間において並列に処理される可能性が高くなり、コマンド処理の効率が向上する。

【００２５】本発明に係る外部記憶装置においては、前記外部記憶制御部は起動時に前記ディスク装置のシーク時間及び回転時間を測定しこの測定値に基づいて前記コマンド処理時間予測値を設定する予測処理時間設定手段を有することを特徴とする。本発明によれば、シーク時間及び回転時間の実測値を利用することにより、コマンド処理時間予測値の精度が高まり、外部記憶制御部によるコマンド処理の効率向上のための制御精度が上がる。

【００２６】

【発明の実施の形態】以下、本発明の好適な実施形態を図面を参照して説明する。

【００２７】［実施形態１］図１は、本発明が適用される共有外部記憶装置の模式的な構成図である。共有外部記憶装置３０は、ホスト計算機３２、３４、３６で共有される外部記憶装置である。共有外部記憶装置３０は、データを格納する３つのディスク装置３８（ディスク装置Ａ、Ｂ、Ｃ）を有し、これらがそれぞれ記憶ディスク部を構成する。また共有外部記憶装置３０は、外部記憶制御部４０を有し、これが各ホスト計算機とディスク装置３８との間のデータの入出力を制御する。ディスク装置３８と外部記憶制御部４０とはＳＣＳＩバス４２により互いに接続されている。なお、各ホスト計算機と外部記憶制御部４０との間もそれぞれＳＣＳＩバス４４により接続されている。

【００２８】図２は共有外部記憶装置３０の内部処理を説明するためのブロック図である。外部記憶制御部４０内の処理を中心に説明する。受入キュー５０が、ホスト計算機３２、３４、３６から発行された読み出しコマンド又は書き込みコマンドを一括して受付て蓄積する。本外部記憶制御部４０は予測処理時間生成手段として、コマンド処理時間予測値を格納した予測値テーブル５２を有している。このコマンド処理時間予測値は、ディスク装置３８の最大シーク時間、最大回転待ち時間、ディスク装置３８内部のデータ転送レート、ＳＣＳＩバス４２のデータ転送レート等の性能値から予測したコマンドの実行時間であり、例えば読み出し／書き込みの別や転送されるデータ量などによって異なる。

【００２９】コマンドバッチ生成部５４は、予測値テーブル５２に基づいて受入キュー５０内のコマンドの実行時間を予測し、コマンド処理時間予測値の和が所定のバッチ時間ｔを超えない範囲でできる限り多くのコマンド５６を受入キュー５０の先頭から読み込む。コマンドバッチ生成部５４は、読み込んだ一群（バッチ）のコマンドの処理が終了するまで、次の読み出し動作を行わない。すなわち、コマンドバッチ生成部５４は、受入キュー５０をバッチ処理する。コマンド５６はコマンドバッチ生成部５４から送出キュー５８に格納される。

【００３０】コマンドキューイング部６０は後述するやり方で送出キュー５８から１つずつコマンドをＳＣＳＩバス４２に送出する。各ディスク装置３８はディスク側キュー６２を有しコマンド・キューイングをサポートしている。よってコマンドキューイング部６０はＳＣＳＩバス４２がフリーであれば、先に送出したコマンドのレスポンスがディスク装置３８から返っていなくても、同一のディスク装置３８に対して多重にコマンドを発行することができる。各ディスク装置３８は、自身に多重に発行されたコマンドをディスク側キュー６２に蓄積し、これから順次読み出して実行する。実行された結果、例えば読み出しデータはディスク装置３８からＳＣＳＩバス４２を介して、外部記憶制御部４０に送られるが、この戻りの経路は図示していない。

【００３１】図３は、コマンドキューイング部６０の処理を説明するための送出キューの模式図である。コマンドバッチ生成部５４は受入キュー５０から上述したようにバッチ時間ｔを基準として１群のコマンドを読み出す。図３に示す例は、１０個のコマンドが読み出され、そのうち４個がディスク装置Ａ、１個がディスク装置Ｂ、５個がディスク装置Ｃを対象とするものであり、これらは送出キュー５８に格納される。キュー７０、７２、７４はそれぞれディスク装置Ａ、Ｂ、Ｃに対応する処理待ちコマンドの列を表している。これらキューに含まれる１～１０の番号を付したブロックは各コマンドを表し、これら各ブロックの高さは各コマンドのコマンド処理時間予測値を表す。よって各キューの縦方向の長さは各記憶ディスク部における処理時間の予測値である。

【００３２】コマンドキューイング部６０は、送出キュー５８から処理時間の予測値が最長の記憶ディスク部に対応するコマンドを１つ選択して読み出し、ＳＣＳＩバスを介して、対応するディスク装置３８に送出する。例えば、図３においてディスク装置Ｃに対応するキュー７４が最長であるので、その中のコマンド（番号“１”で表す）を送出する。この状態で、まだ同じキュー７４が最長なので、このキュー７４からコマンド“２”を送出する。すると今度は、ディスク装置Ａに対応するキュー７０が最長となるので、そのキュー７０からコマンド“３”を送出する。同様に続いてコマンド“４”～“１０”が順次送出される。なお、これら各キュー間の比較はコマンド処理時間予測値によるものであり、コマンド数には依存しない。例えば、コマンド“５”まで送出した時点で、キュー７０、７４にはコマンドがそれぞれ２

個残っており、キュー７２には１個しか残っていない。しかし処理時間はキュー７２の方が他の２つのキューよりも長いので、他のキューより優先してキュー７２のコマンド“６”が送出される。

【００３３】図４、図５は上述したコマンドキューイングの効果を説明するための図である。図４は、ディスク装置３８を２つとした場合における送出キュー内のコマンド列の１例を示す模式図である。図５は、図４に示した例における処理のタイミングチャートである。図４に示す例では、ディスク装置Ａに対応するキュー８０は３個のコマンドａ、ｂ、ｃを有し、ディスク装置Ｂに対応するキュー８２は１個のコマンドを有する。コマンドａ、ｂ、ｃ、ｄはそれぞれコマンド処理時間予測値として２τ、４τ、６τ、７τを有する（τは任意の時間単位）。なお、コマンドをディスク装置３８に送る際、及びディスク装置３８からコマンドの処理結果を返す際の、ＳＣＳＩバスの使用権の獲得からその開放までの時間をそれぞれセンド時間、リターン時間と称することとする。センド時間、リターン時間はこの例ではτである。図５（ａ）は本実施形態のコマンドキューイング部の処理のタイミングチャートを示す。一方、図５（ｂ）は、本実施形態のコマンドキューイング部とは異なる処理のタイミングチャートを示す。コマンドの対応するディスク装置への送出の順序は、図５（ａ）に示す本実施形態の場合においてはコマンドｃ、ｄ、ｂ、ａ（それぞれセンド時間９０、９２、９４、９６に対応）の順であり、一方、図５（ｂ）に示す例ではコマンドｄ、ｃ、ｂ、ａ（それぞれセンド時間１００、１０２、１０４、１０６に対応）の順である。送出されたコマンドはディスク装置において実行された後、リターン時間１１０、１１２、１１４、１１６、１２０、１２２、１２４、１２６において、外部記憶制御部４０に応答が返される。

【００３４】本実施形態では、２つのキューのうち短い処理時間を有するキュー８２の処理期間（コマンドｄの処理期間）は処理時間の長い方のキュー８０の処理期間（コマンドｃ、ｂ、ａの処理期間）に包含される。このように、上記コマンドキューイング部の処理によれば、短い処理時間を有する記憶ディスク部の処理期間は、それより長い処理時間を有する記憶ディスク部の処理期間の背後に隠されるように、処理がスケジューリングされる。これに対し、本実施形態を異なる図５（ｂ）に示す方法では、短い処理時間を有する記憶ディスク部の処理期間は、それより長い処理時間を有する記憶ディスク部の処理期間からはみ出す。図５（ａ）、（ｂ）に示す例では処理開始から終了までの時間は、それぞれ１４τ、１５τである。このように本発明の実施形態においては、全記憶ディスク部に対するコマンド処理終了までの時間が短縮され、レスポンス及びスループットが向上する。

【００３５】本実施形態の他の特徴を次に説明する。コマンドの処理に際して、上述のようにＳＣＳＩバスの使用に関するセンド時間、リターン時間が必要になる。これらの時間中には、外部記憶制御部４０やディスク装置３８がバス４２の使用権を獲得するアービトレーションという動作を実行する必要がある。しかしディスク装置３８がバス・フリー状態を検出してアービトレーションを開始するまでの時間は、外部記憶制御部４０のその時間よりも遅い。さらに、ディスク装置３８のＳＣＳＩＩＤの優先順位は、外部記憶制御部４０のそれよりも低い値に設定されている。そのため、外部記憶制御部４０が連続的にコマンドをディスク装置３８のいずれかに送出しようとする場合、ディスク装置３８側は、バス４２の使用権を獲得することが困難となり、ディスク装置３８からのレスポンス時間が長くなる可能性がある。

【００３６】図６、図７は、本装置の外部記憶制御部４０における上記問題を解決するための上記方法を説明するフローチャートである。図６は、本装置の電源オン時における自己診断での動作を説明するフローチャートである。外部記憶制御部４０は、本装置の電源オン時の自己診断の一環として、バス・フリー状態を検出し試行的にコマンドを発行して、ディスク装置３８がこのコマンドに対しレスポンスを返すためにバス４２の使用権を獲得するまでの時間を測定する。ここで発行されるコマンドは、ディスクコネクトを要求するがディスク装置３８内のデータに影響を及ぼさないコマンドであり、例えばＩＮＱＵＩＲＹ、ＲＥＡＤ、ＣＡＰＡＣＩＴＹ等である。外部記憶制御部４０はこれらのコマンドを複数回発行して、それらに対してディスク装置３８からのレスポンスを得るまでの応答時間の平均値を計測する。この応答時間の平均値は、コマンドキューイング部６０に与えられる。図７は実際のコマンド発行処理におけるコマンドキューイング部６０の動作を説明するフローチャートである。コマンドキューイング部６０はコマンドに対するアービトレーションの開始をバス・フリー状態の検出から前記応答時間の平均値に応じた時間、例えば、平均値に安全係数を乗じた時間だけ遅延させる。ちなみに、安全係数はデフォルト値として例えば２を与える。なお、アービトレーションの実行を遅延させている間にバス・ビジー状態となった場合には、バス・フリー状態となるのを待って再度処理を試みる処理を行う。

【００３７】上記処理により、ディスク装置３８のバス獲得における優先度が増し、外部記憶制御部４０がディスク装置３８からのレスポンスを迅速に受け取ることができる。また、遅延時間はディスク装置３８ごとに変えることができ、バス・フリー状態の検出からアービトレーションの開始までの時間がディスク装置３８ごとに異なる場合でも対応可能である。

【００３８】また本装置では、複数の読み出しコマンドに対するホスト計算機３２へのデータ転送処理を一括して行う。これによるＳＣＳＩバス４４を介したホスト計

算機３２とのデータ転送の処理時間は、その処理をコマンドごとに行う場合より短くなり、ホスト計算機３２に対するレスポンスが向上する。この一括転送処理を以下に説明する。外部記憶制御部４０は、１回のコマンドのバッチ中に含まれる読み出しコマンドのうち処理待ち状態にあるコマンド数を計数する機能と、ディスク装置３８からの読み出しデータを蓄積することができるバッファとを有する。外部記憶制御部４０は、バッファにできるだけ多くの読み出しコマンドに対する読み出しデータを蓄積し、バッファが一杯になると、バッファからホスト計算機３２への転送処理を開始する。これにより、ホスト計算機３２とのバスの接続処理といったオーバーヘッドが減るのでホスト計算機３２に対するレスポンスが向上する。前記処理待ちの読み出しコマンドの数が０になれば、バッファが一杯にならなくても、外部記憶制御部４０はバッファ内のデータをホスト計算機３２に転送する。また、外部記憶制御部４０は、コマンド処理時間予測値に基づいてコマンド処理の予測スケジュールを作成することができるので、このスケジュールに基づいて読み出しコマンドに対するディスク装置からの応答の間隔を予測することができる。外部記憶制御部４０は、その間隔とホスト計算機に対する転送処理のオーバーヘッド時間との比較に基づいて、バッファに次の読み出しコマンドのデータが到着するのを待つことの得失を判断して、バッファからの転送処理を制御してもよい。

【００３９】本装置は、なんらかのエラーにより発行されたコマンドに対するレスポンスがない場合のタイムアウト処理を以下のように行う。ディスク装置３８は、そのディスク側キュー６２内のコマンドを例えばアクセスされるアドレスに基づいて最適な順序で実行する。外部記憶制御部４０は、ディスク装置３８側にキューイング済みであり処理待ちであるコマンドのコマンド処理時間予測値の最大値をタイムアウト値として保持するとともに、直前のコマンドの処理終了からの経過時間を計測する。

【００４０】図８はタイムアウト処理を説明する模式図である。図８（ａ）はある１つのディスク装置３８に４つのコマンドＡ、Ｂ、Ｃ、Ｄを発行した場合の、コマンド処理時間予測値によるディスク装置の占有時間のタイムチャートである。コマンド処理時間予測値の大小関係は、Ａ＜Ｃ＜Ｄ＜Ｂであるとする。外部記憶制御部４０は、ディスク装置において最初に処理されるコマンドＡの処理が完了し、そのレスポンスが返された時点から経過時間を計測し始める。またこれと同時にこのディスク装置に対するタイムアウト値は、処理待ちコマンドＢ、Ｃ、Ｄのうち最長のコマンド処理時間予測値を有するコマンドＢのその予測値を設定される（図８（ｂ））。次にコマンドＢの処理が開始される。ディスク装置はコマンドＢの処理をタイムアウト値以内に完了し、外部記憶制御部４０にレスポンスを返す（図８（ｃ））。外部記憶制御部４０はこの時点においてはディスク装置３８が正常動作を行っていると判断し、経過時間をリセットし、またタイムアウト値をコマンドＤのコマンド処理時間予測値に変更する（図８（ｄ））。

【００４１】図８（ａ）に示すコマンド実行のスケジュールはディスク装置により決定され、外部記憶制御部４０はその実行順序を知らない。よって、外部記憶制御部４０はコマンドＢのレスポンスがあった時点で、次にコマンドＣ、Ｄのいずれが実行されるかを知らない。しかし、タイムアウト値は、まだ実行されていないコマンドのうち最長のコマンド処理時間予測値である。よって、外部記憶制御部４０は新たに設定されたタイムアウト値以内になんらのレスポンスもディスク装置から得られなければ、コマンドＣ、Ｄのいずれも実行されずエラーと判断することができる（図８（ｅ））。

【００４２】このエラーが検出されるまでの時間は、コマンドＡ、Ｂの処理に要した実時間とコマンドＤのコマンド処理時間予測値との和である。もしコマンドのバッチ単位でタイムアウトを監視したならば、エラーはディスク装置の占有時間（コマンドＡ、Ｂ、Ｃ、Ｄのコマンド処理時間予測値の和）が経過しないと検出されない。よって上述のようにコマンド単位でタイムアウトを監視することにより、タイムアウトの検出が早まる。これによりエラーに対する対応がそれだけ早く行われ、外部記憶装置のレスポンスの向上を図ることができる。

【００４３】さらに、外部記憶制御部４０は、バッチ（コマンドのグループ）の処理が最適に行われるように、コマンド発行のスケジューリングを行う。コマンド・キューイング機能を有するディスク装置３８は、それ自身がスケジューリング機能を有している。このディスク装置３８におけるスケジューリングは、ディスク装置の回転待ち時間、ヘッドのシーク時間、不良セクタの置き換えなどを考慮して行われる。これらの点は外部記憶制御部４０では把握困難であるので、これらの点に基づくスケジューリングはディスク装置側で行うのがよい。

【００４４】しかし、ディスク装置側は、そのディスク側キュー６２に蓄積されたコマンドの範囲内でスケジューリングを行い、外部記憶制御部４０から未発行でディスク側キュー６２に無いコマンドをスケジューリングに際して考慮することができないので、バッチ単位で見た場合、必ずしも最適な順序で処理が行われていない可能性がある。例えば、ディスク側キュー６２が空であるときに受け取ったコマンドは、無条件で実行開始されるため、シーク時間や回転待ち時間が多くかかるコマンドから処理が行われる可能性がある。

【００４５】そこで、外部記憶制御部４０は送出キュー５８内のコマンドを各ディスク装置ごとにそのアクセスする論理アドレスに基づき並べ換えて、順に各ディスク装置に発行する。このとき、コマンド処理時間の予測値が最長であるディスク装置に対するコマンドを取り出す

ことは上述した通りである。また、バッチが連続する場合には、後のバッチの処理は、アドレスに基づいて前のバッチの最後に実行したコマンドに対して実行時間が短いと予想されるコマンドから始める。具体的には、コマンドの発行順序をアクセスする論理アドレスの大きい順もしくは小さい順に発行する。また例えば、前のバッチが論理アドレスの大きい順にコマンドを発行した場合は、それに続くバッチは論理アドレスの小さい順にコマンドを発行する。このように外部記憶制御部４０が論理アドレスに基づいてアクセス時間が最適となるようにコマンドを発行することにより、ディスク装置３８がディスク側キュー６２にのみ基づいてスケジューリングを行う場合の上記問題が解決され、外部記憶装置のレスポンスが向上する。

【００４６】［実施形態２］図９は、本発明が適用される他の共有外部記憶装置の模式的な構成図である。図中、特に断らない限り、図１の符号に２００を加えた構成要素は、図１と同一の機能を有するものとする。本実施形態と実施形態１との主たる相違点は、本実施形態の外部記憶装置２３０は２つの外部記憶制御部２５０、２５２を有し、これらがディスク装置２３８を共有して制御する点にある。外部記憶制御部２５０、２５２及び２つのディスク装置２３８（ディスク装置Ａ、Ｂ）はＳＣＳＩバス２５４にて接続される。両外部記憶制御部間には、それらの間での通信を可能にするバス２５６が設けられている。両外部記憶制御部は、それぞれ実施形態１の外部記憶制御部４０と基本的には同一の動作を行う。外部記憶制御部２５０、２５２が、外部記憶制御部４０と異なるのは、一方の外部記憶制御部からバッチを処理した後、続いて他のバッチをもう一方の外部記憶制御部から処理する場合である。これは、外部記憶制御部を２つ有する本装置に特有である。

【００４７】本装置では、一方の外部記憶制御部が処理するバッチにおける最後のコマンドのアクセス先の論理アドレスが、ディスク装置のアドレスの最小値、最大値のいずれに近いかを、バス２５６を介して他方の外部記憶制御部に通知する。すなわち、バス２５６はアドレス情報通知手段として機能する。例えば、外部記憶制御部２５０は各ディスク装置２３８ごとに、その最後のコマンドの論理アドレスが０に近い場合には、バス２５６にＬ（ｌｏｗ）レベルの制御信号を送出し、逆に論理アドレスが最大値に近い場合にはＨ（ｈｉｇｈ）レベルの制御信号を送出する。例えば、外部記憶制御部２５.２はバス２５６からディスク装置Ａについ てＬレベル、ディスク装置Ｂについ てＨレベルの制御信号を受けた場合には、そのディスクＡに対しては、そのアクセスする論理アドレスの小さいコマンドから順に発行し、ディスクＢに対しては、そのアクセスする論理アドレスの大きいコマンドから順に発行する。

【００４８】このように２つの外部記憶制御部２５０、２５２間においても、コマンドが論理アドレスに基づいてアクセス時間が最適となるよう発行され外部記憶装置のレスポンスが向上する。

【００４９】［実施形態３］図１０は、レベル１のＲＡＩＤディスク・アレイを備え本発明が適用される共有外部記憶装置の模式的な構成図である。図中、特に断らない限り、図９の符号に１００を加えた構成要素は、図９と同一の機能を有するものとする。本実施形態と実施形態２との主たる相違点は、本実施形態の外部記憶装置３３０は２つのレベル１のＲＡＩＤディスク・アレイ（ＲＡＩＤ１ディスク・アレイ）３５８、３６０を記憶ディスク部として有する点にある。各ＲＡＩＤ１ディスク・アレイは２つのディスク装置から構成され、これらディスク装置間でミラーリングが行われる。各ＲＡＩＤ１ディスク・アレイを構成する２つのディスク装置は互いに異なるＳＣＳＩバス３６２、３６４に接続される。これら２つのＳＣＳＩバス３６２、３６４は、外部記憶制御部３５０、３５２にもそれぞれ接続される。両外部記憶制御部３５０、３５２は基本的には、各ＲＡＩＤ１ディスク・アレイをそれぞれ１個の記憶ディスク部として扱い、実施形態２と同様の制御を行う。但し、本装置においては、読み出し処理における最適化を行う点が、上記実施形態と異なる。以下、この点を、ＲＡＩＤ１ディスク・アレイ３６０を例にとって説明する。

【００５０】ＲＡＩＤ１ディスク・アレイ３６０では、書き込み処理は、同一のデータを２つのディスク装置３６６、３６８の双方に書き込まなければならない。しかし、読み出し処理については、ディスク装置３６６、３６８のいずれか一方から読み出せばよい。そこで、外部記憶制御部３５０、３５２は、コマンド処理時間予測値に基づいて、ＲＡＩＤ１ディスク・アレイ３６０に対する読み出し処理をディスク装置３６６、３６８に振り分け、並列処理させる。これにより、コマンドの全体の処理時間が短縮され、装置のレスポンスが向上する。

【００５１】表１は、上記処理を説明するためのＲＡＩＤ１ディスク・アレイ３６０に対するコマンド群の例を表しており、これらは外部記憶制御部３５０から発行されるものとする。表にはコマンドの読み出し／書き込みの種別とコマンド処理時間予測値である予測実行時間が示されている。

【００５２】

【表１】

	コマンド	予測実行時間
a	READ	３０ｍｓ
b	READ	１２０ｍｓ
c	WRITE	４０ｍｓ
d	READ	４０ｍｓ
e	WRITE	３５ｍｓ
f	READ	５０ｍｓ

コマンドｃ、ｅは書き込みコマンドであるので、ディスク装置３６６、３６８の双方に発行される。残りのコマンドａ、ｂ、ｄ、ｆは読み出しコマンドであるので、ディスク装置３６６、３６８のいずれかに発行されればよい。外部記憶制御部３５０は、これら読み出しコマンドをそのコマンド処理時間予測値の和がなるべく均等化するように２つのグループに分ける。ここでは、コマンドａ、ｄ、ｆとコマンドｂとに分ける。これらは各グループは１２０ｍＳのコマンド処理時間予測値の和を有する。この外部記憶制御部３５０は、コマンドａ、ｄ、ｆをディスク装置３６６に発行し、コマンドｂをディスク装置３６８に発行する。

【００５３】上記の処理はＲＡＩＤタスクにて行われる。ＲＡＩＤタスクはコマンドバッチ生成手段であるモニタにより生成されたコマンド群を受け取り、それを上記の手法により効率的にディスク装置に割り振る。

【００５４】［実施形態４］図１１は、レベル５のＲＡＩＤディスク・アレイを備え本発明が適用される共有外部記憶装置の模式的な構成図である。図中、特に断らない限り、図１０の符号に１００を加えた構成要素は、図１０と同一の機能を有するものとする。本実施形態と実施形態３との主たる相違点は、本実施形態の外部記憶装置４３０は２つのレベル５のＲＡＩＤディスク・アレイ（ＲＡＩＤ５ディスク・アレイ）４５８、４６０を記憶ディスク部として有する点にある。各ＲＡＩＤ５ディスク・アレイは５つのディスク装置から構成され、これらディスク装置の格納するデータから生成されるパリティは、ディスク装置間で分割して保持される。各ＲＡＩＤ５ディスク・アレイを構成する５つのディスク装置は互いに異なるＳＣＳＩバス４６２、４６４、４６６、４６８、４７０に接続される。これら５つのＳＣＳＩバスは、外部記憶制御部４５０、４５２にもそれぞれ接続される。両外部記憶制御部４５０、４５２は基本的には、各ＲＡＩＤ５ディスク・アレイをそれぞれ１個の記憶ディスク部として扱い、実施形態２と同様の制御を行う。但し、本装置においては、パリティの書き込み処理における最適化を行う点が、上記実施形態と異なる。以下、この点を、ＲＡＩＤ５ディスク・アレイ４６０を例にとってを説明する。

【００５５】ＲＡＩＤ５ディスク・アレイ４６０では、データの書き込み処理はパリティの更新処理を伴う。例えばｎ個のデータをＤｋ（ｋ＝１～ｎ）、これらにより生成されるパリティをＰとする。これらはレベル５のＲＡＩＤディスク・アレイを構成する各ディスク装置に分散して保持されている。ここでＤｊ（１≦ｊ≦ｎ）をＤｊ’で更新する書き込み処理を考える。新たなｎ個のデータに対するパリティを求める第１の方法はＤｊ’、Ｄｊ及びＰから求める方法であり、第２の方法はＤｊ’及びＤｋ（ｋ≠ｊ）から求める方法である。この第１の方法においてはＤｊ及びＰをディスク装置から読み出さなければならない。一方、第２の方法においてはＤｋ（ｋ≠ｊ）を読み出さなければならない。この第１の方法と第２の方法とでは、読み出し処理が行われるディスク装置が異なる。そこで外部記憶制御部４５０、４５２は、ＲＡＩＤ５ディスク・アレイ４６０に対する複数の書き込みコマンドについての上記パリティ更新における読み出し処理が各ディスク装置に分散し処理時間が均等化するように、上記第１の方法、第２の方法を選択する。すなわち書き込みコマンドに伴う読み出し処理が各ディスク装置間において並列に処理される。これにより、コマンドの全体の処理時間が短縮され、装置のレスポンスが向上する。

【００５６】表２は、上記処理を説明するためのＲＡＩＤ５ディスク・アレイ４６０に対するコマンド群の例を表しており、これらは外部記憶制御部４５０から発行されるものとする。表にはコマンドの読み出し／書き込みの種別、アクセス先の論理アドレス、データ長及び予測実行時間が示されている。ここで書き込みコマンドの予測実行時間は、パリティ計算のための読み出し動作に必要な時間を含まないコマンド処理時間予測値である。

【００５７】

【表２】

	コマンド	論理アドレス	データ長(ﾌﾞﾛｯｸ数)	予想実行時間
a	READ	２５４３	１	３１ｍｓ
b	WRITE	３２４	１	６２ｍｓ
c	WRITE	６８６	１	６２ｍｓ
d	READ	２１	１	３１ｍｓ
e	WRITE	７６０	１	６２ｍｓ
f	WRITE	５８９０	１	６２ｍｓ

コマンド処理時間予測値を求める際に、ディスク装置４７２、４７４、４７６、４７８、４８０（ディスク装置Ａ、Ｂ、Ｃ、Ｄ、Ｅ）に関して、最大シーク時間は１８ｍＳ、最大回転待ち時間は１２ｍＳ、内部データ転送レートは５ＭＢ／Ｓ及びコントローラのオーバーヘッドは０．７ｍＳであるという条件を使用し、また１ブロックのデータ長は１ＫＢ、ＳＣＳＩバス４６２、４６４、４６６、４６８、４７０の転送レートは１０ＭＢ／Ｓであるとした。

【００５８】図１２はＲＡＩＤ５ディスク・アレイ４６０のセクタ・アドレスの割付けマップを表す模式図である。セクタ・アドレスは図１２に示されるような規則性

に基づいて割り付けられる。この割付け方と表２の各コマンドのアドレスとに基づいて、各ディスク装置ごとの処理時間の予測値を求めると、ディスク装置Ａ、Ｂ、Ｃ、Ｄ、Ｅそれぞれについて、９３ｍＳ、９３ｍＳ、０ｍＳ、６２ｍＳ、６２ｍＳとなる。

【００５９】さて、コマンドｂ、ｃ、ｅ、ｆは書き込みコマンドであるので、上述のようにパリティ計算のため、ディスク装置に対する読み出し動作を伴う。例えば、コマンドｂに対する上述の第１のパリティ生成方法は、ディスク装置Ｂの古いデータとディスク装置Ａの古いパリティとを読み出す動作を伴い、また第２のパリティ生成方法は、ディスク装置Ｃ、Ｄ、Ｅの対応セクタの古いデータを読み出す動作を伴う。パリティ計算のための読み出し動作を含まないディスク装置の占有時間（表２）に基づく比較では、ディスク装置Ａ、Ｂの負荷が高くなっている。そこでコマンドｂのパリティ生成処理は上記第２の方法により行い、ディスク装置間でその占有時間を平均化する。コマンドｃ、ｅ、ｆについても上記第１、第２の方法のいずれかを、ディスク装置間の処理の均等化が図られるように選択する。これによりコマンド全体の処理が効率化され、外部記憶装置４３０のレスポンスが向上する。

【００６０】上記の処理はＲＡＩＤタスクにて行われる。ＲＡＩＤタスクはコマンドバッチ生成手段であるモニタにより生成されたコマンド群を受け取り、それを上記の手法により効率的にディスク装置に割り振る。

【００６１】また、上では、コマンド処理時間予測値を、接続されるディスク装置の最大シーク時間、最大回転待ち時間等を基に算出しているが、これらの値に代えて、予め測定により得た評価値を用いて算出することもできる。測定値を取り入れることで、コマンド処理時間予測値の精度が向上し、外部記憶制御部の制御精度が向上する。これにより、装置のレスポンスの向上も期待できる。

【００６２】具体的には、外部記憶制御部４５０、４５２は電源オン時に自己診断として、ＳＥＥＫコマンドを複数回発行する。そして、その際にシーク・回転時間を測定し、それぞれの平均値を算出し、その値に安全係数を掛けた値をシーク時間、回転待ち時間の評価値とする。この評価値に基づきコマンドの実行時間を予測することにより、最大値を使用した場合より精度のよい予測が可能となり、処理の効率向上が図られる。なお安全係数は２をデフォルト値とし、また、システムごとに選択・変更可能とする。

【００６３】［実施形態５］模式的な構成は図１１と同一の装置であって記憶ディスク部４５８、４６０をレベル３のＲＡＩＤディスク・アレイで置換した装置にも本発明は適用される。この場合を図１１を流用して、実施形態４と異なる点のみを説明する。本実施形態と実施形態４との主たる相違点は、記憶ディスク部内のデータの持ち方にある。図１３は、従来のＲＡＩＤ３ディスク・アレイのセクタ・アドレスの割付けマップを表す模式図である。従来においては、１つのディスク装置（図ではディスク装置Ｅ）がパリティ専用に割り当てられていた。これに対し、図１４は本装置におけるＲＡＩＤ３ディスク・アレイのセクタ・アドレスの割付けマップを表す模式図である。本装置では、パリティを格納するディスク装置は、１２８セクタごとに変わる。すなわち、パリティは、０～１２７セクタ（０～５１１バイトに相当）ではディスク装置Ｅに格納され、次の１２８～２５５セクタ（５１２～１０２３バイト）ではディスク装置Ａに格納され、以降１２８セクタごとにディスク装置Ｂ、Ｃ、Ｄ、…に格納される。

【００６４】ここで、外部記憶制御装置が５ブロック分（０～２５６０バイト）のデータを読み出すコマンドを発行した場合を説明する。既存のＲＡＩＤ３ディスク・アレイでは、データが格納されている４つのディスク装置Ａ～Ｄのそれぞれに６４０回のリード・アクセスを行う必要があった。ところが、本装置では、セクタ単位でパリティが割り振られているため、５つのディスク装置Ａ～Ｅのそれぞれに５１２回のリード・アクセスを行うこととなる。これにより、アクセスがディスク装置間にて均等化され、２０％性能が向上する。この方式は、１度に大量のデータ転送を必要とする画像データ等のマルチメディアアプリケーション用データの入出力において特に効果を発揮する。

【００６５】

【発明の効果】本発明の外部記憶装置によれば、複数の記憶ディスク部を有する外部記憶装置において、装置全体のレスポンスやスループットが優れているとともに、記憶ディスク部間におけるレスポンスやスループットの不均衡がないという効果がある。コマンドの処理をバッチごとに進めることにより最大レスポンス時間、最低スループットが保証された外部記憶装置が得られるという効果がある。

【００６６】また、本発明の外部記憶装置によれば、バス獲得における優先度に関し、外部記憶制御部と記憶ディスク部との間の均衡がとれ、記憶ディスク部からのレスポンス時間が短縮され、装置のレスポンスが向上するという効果がある。

【００６７】本発明の外部記憶装置によれば、複数の読み出しコマンドに対応するデータが一括してコンピュータへ転送されるので、その転送の際の外部記憶装置からコンピュータへのチャネルの接続処理といった時間的オーバーヘッドが低減され、レスポンスやスループットが向上するという効果が得られる。

【００６８】本発明の外部記憶装置によれば、コマンド単位でタイムアウトを監視することにより、タイムアウトの検出が早まる。これによりエラーに対する対応がそれだけ早く行われ、外部記憶装置のレスポンスの向上を

図ることができる。

【００６９】本発明の外部記憶装置によれば、送出キュー内の前記コマンドにより指定される前記アドレスが例えば昇順、降順となるようにコマンドが選択され、これらコマンドの実行における記憶ディスク部のヘッドのシーク時間やディスクの回転時間といったアクセス時間が最小となる。これによりレスポンスやスループットが向上するという効果が得られる。

【００７０】本発明の外部記憶装置によれば、複数の外部記憶制御部を有する場合、一方の外部記憶制御部において最後にアクセスされたアドレスに関する情報を、次の処理を行う外部記憶制御部に通知する。次の処理を行う外部記憶制御部はこのアドレス情報に基づいて、例えば前記アクセス時間が最小であるコマンドから処理を開始するので、レスポンスやスループットが向上するという効果が得られる。

【００７１】本発明の外部記憶装置によれば、送出キュー中に複数の読み出しコマンドがあれば、それらをＲＡＩＤ１アレイ・ディスクを構成する２つのディスク装置に振り分けて発行し、これら両ディスク装置において異なる読み出しコマンドを並列に処理させることにより、コマンド処理の効率が向上する。

【００７２】本発明の外部記憶装置によれば、ＲＡＩＤ５アレイ・ディスクを記憶ディスク部とする場合、送出キュー中に複数の書き込みコマンドがあれば、それらに伴う上記パリティ更新における読み出し処理が、各ディスク装置に分散し処理時間が均等化するように、上記第１の方法、第２の方法を選択する。これにより、書き込みコマンドに伴う読み出し処理が各ディスク装置間において並列に処理され、コマンド処理の効率が向上するという効果が得られる。

【００７３】本発明の外部記憶装置によれば、ＲＡＩＤ３アレイ・ディスクを記憶ディスク部とする場合、データの分散度合いが高まり、コマンド処理が各ディスク装置間において並列に処理される可能性が高くなり、コマンド処理の効率が向上するという効果が得られる。

【００７４】本発明の外部記憶装置によれば、シーク時間及び回転時間の実測値を利用することにより、コマンド処理時間予測値の精度が高まり、外部記憶制御部によるコマンド処理の効率向上のための制御精度が上がるという効果がある。

【図面の簡単な説明】

【図１】　本発明が適用される共有外部記憶装置の模式的な構成図。

【図２】　共有外部記憶装置の内部処理を説明するためのブロック図。

【図３】　コマンドキューイング部の処理を説明するための送出キュー内のコマンド列の模式図。

【図４】　コマンドキューイング処理を説明するための送出キュー内のコマンド列の１例を示す模式図。

【図５】　コマンドキューイング処理を説明するタイミングチャート。

【図６】　電源オン時における自己診断の動作を説明するフローチャート。

【図７】　実際のコマンド発行処理におけるコマンドキューイング部の動作を説明するフローチャート。

【図８】　タイムアウト処理を説明する模式図。

【図９】　本発明が適用される他の共有外部記憶装置の模式的な構成図。

【図１０】　ＲＡＩＤ１ディスク・アレイを備え本発明が適用される共有外部記憶装置の模式的な構成図。

【図１１】　ＲＡＩＤ５ディスク・アレイを備え本発明が適用される共有外部記憶装置の模式的な構成図。

【図１２】　ＲＡＩＤ５ディスク・アレイのセクタ・アドレスの割付けマップを表す模式図。

【図１３】　従来のＲＡＩＤ３ディスク・アレイのセクタ・アドレスの割付けマップを表す模式図。

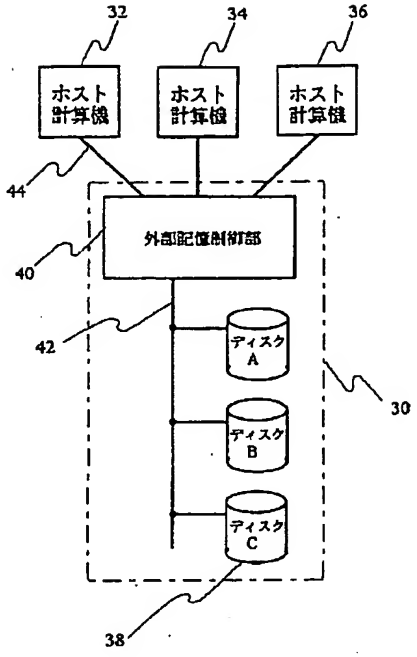
【図１４】　実施形態におけるＲＡＩＤ３ディスク・アレイのセクタ・アドレスの割付けマップを表す模式図。

【図１５】　従来の外部記憶装置の模式的な構成図。

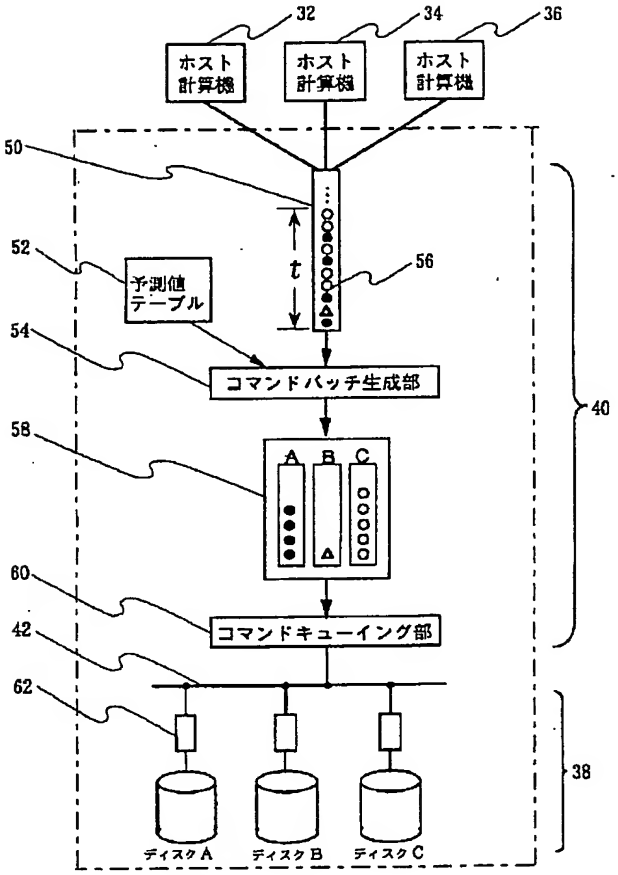
【符号の説明】

３０　共有外部記憶装置、３２，３４，３６　ホスト計算機、３８　ディスク装置、４０　外部記憶制御部、４２，４４　ＳＣＳＩバス、５０　受入キュー、５２　予測値テーブル、５４　コマンドバッチ生成部、５８　送出キュー、６０コマンドキューイング部、６２　ディスク側キュー、２５０，２５２　外部記憶制御部、２５６　バス、３５８，３６０　ＲＡＩＤ１ディスク・アレイ、４５８，４６０　ＲＡＩＤ５ディスク・アレイ。

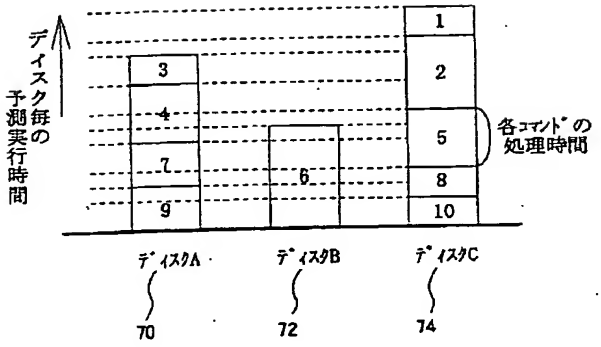
【図１】


32
34
36
ホスト計算機
ホスト計算機
ホスト計算機
44
外部記憶制御部
40
42
ディスクA
ディスクB
ディスクC
30
38

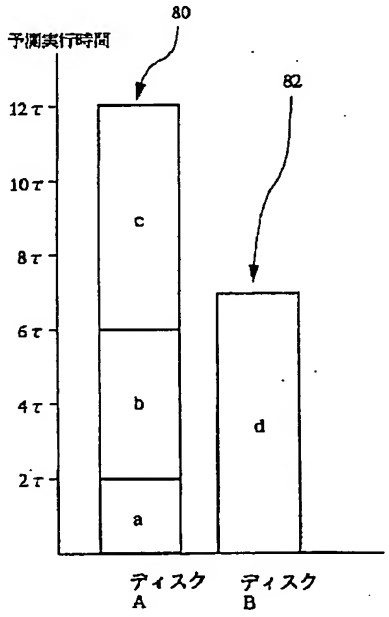
【図２】


32
34
36
ホスト計算機
ホスト計算機
ホスト計算機
50
52
予測値テーブル
t
56
54
コマンドバッチ生成部
40
58
A
B
C
60
コマンドキューイング部
42
62
38
ディスクA
ディスクB
ディスクC

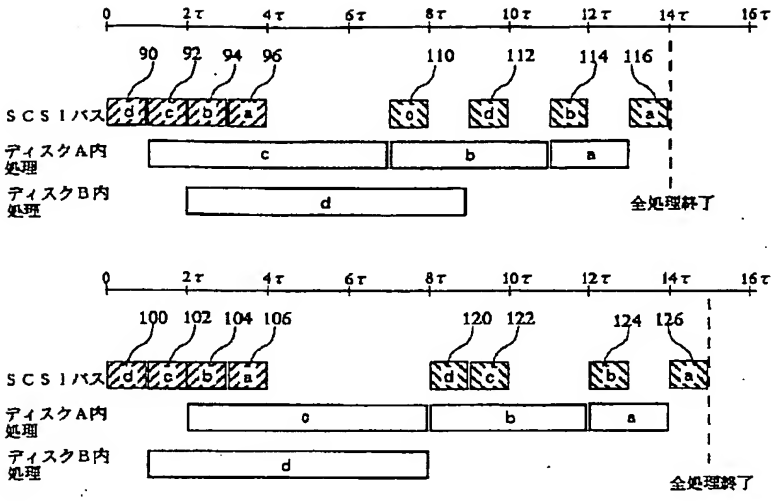
【図３】


ディスク毎の予測実行時間
1
2
3
4
5
6
7
8
9
10
各コマンドの処理時間
ディスクA
ディスクB
ディスクC
70
72
74

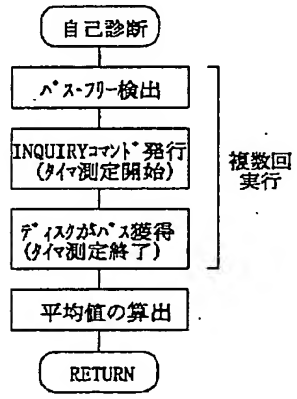
【図４】


80
予測実行時間
82
12τ
10τ
8τ
6τ
4τ
2τ
c
b
a
d
ディスクA
ディスクB

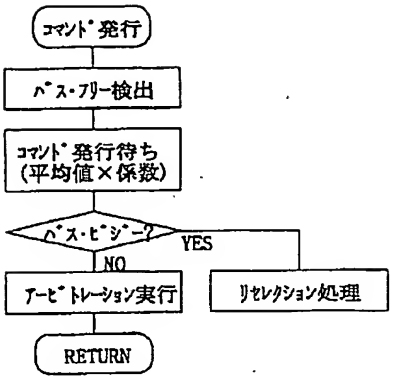
【図５】



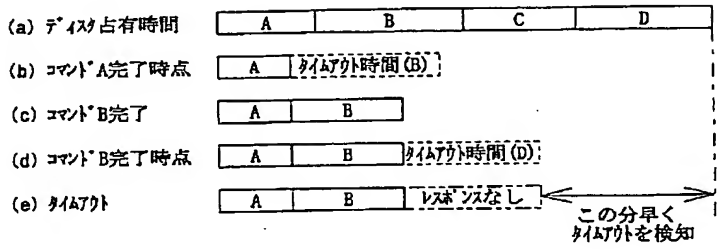
【図６】



【図７】



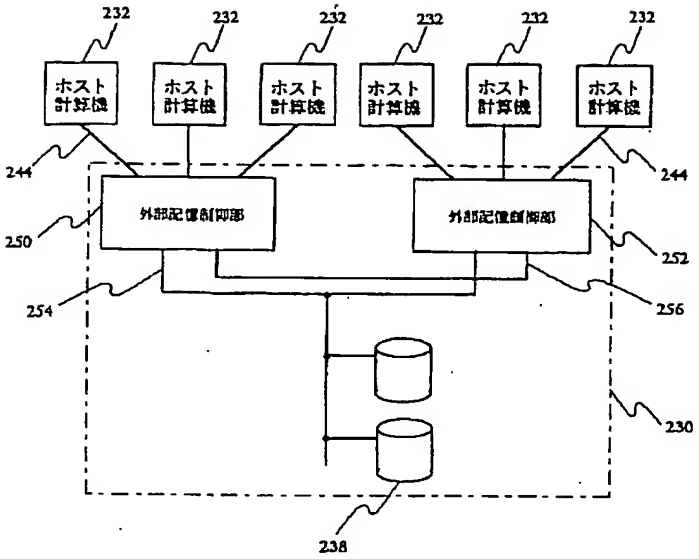
【図８】



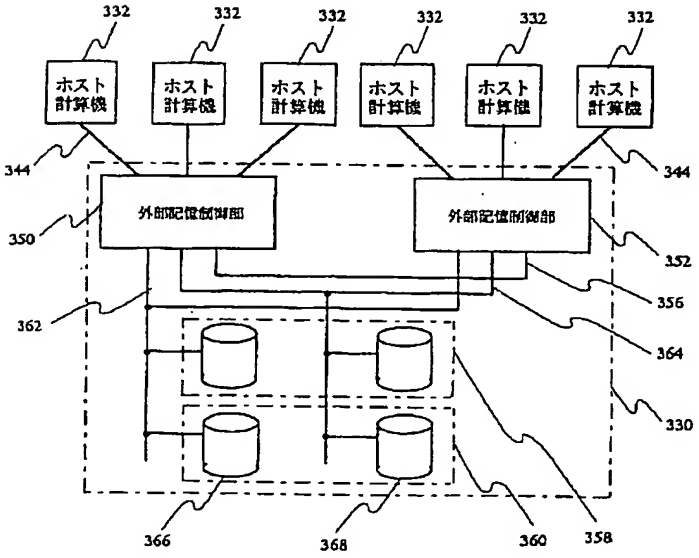
【図１２】

DISK A	DISK B	DISK C	DISK D	DISK E
0	1	2	3	P
P	4	5	6	7
11	P	8	9	10
14	15	P	12	13
17	18	19	P	16
20	21	22	23	P
・	・	・	・	・
・	・	・	・	・
320	321	322	323	P
P	324	325	326	327
・	・	・	・	・
・	・	・	・	・
・	・	・	・	・

【図９】



【図１０】



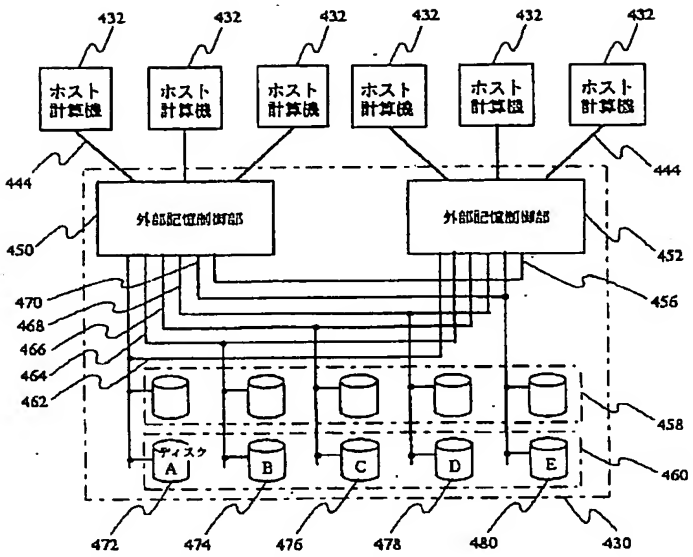
【図１３】

DISK A	DISK B	DISK C	DISK D	DISK E
0	1	2	3	P
508	509	510	511	P
512	513	514	515	P
1020	1021	1022	1023	P
1024	1025	1026	1027	P
1532	1533	1534	1535	P
1536	1537	1538	1539	P
2044	2045	2046	2047	P
2048	2049	2050	2051	P
2556	2557	2558	2559	P
2560	2561	2562	2563	P

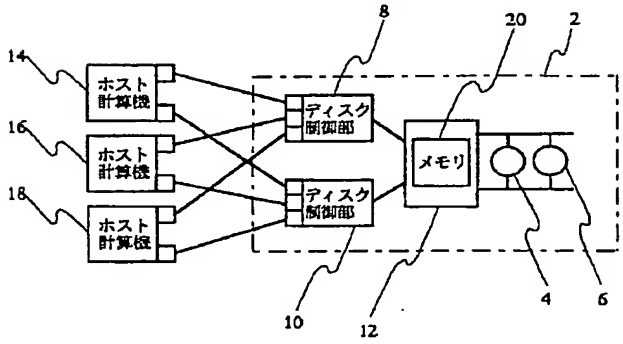
【図１４】

DISK A	DISK B	DISK C	DISK D	DISK E
0	1	2	3	P
508	509	510	511	P
P	512	513	514	515
P	1020	1021	1022	1023
1027	P	1024	1026	1026
1535	P	1532	1533	1534
1538	1539	P	1536	1537
2046	2047	P	2044	2045
2049	2050	2051	P	2048
2557	2558	2559	P	2556
2560	2561	2562	2563	P

【図１１】


432
ホスト計算機
444
450
外部記憶制御部
452
456
470
468
466
464
462
458
460
ディスク
A
B
C
D
E
472
474
476
478
480
430

【図１５】


8
20
2
14
16
18
ホスト計算機
ディスク制御部
メモリ
10
12
4
6